HUAWEI CLOUD

27.03.2024 – 20:38 Uhr

Huawei Cloud auf der KubeCon EU 2024: Entfesselung des intelligenten Zeitalters mit kontinuierlicher Open-Source-Innovation

Paris, 27. März 2024 (ots/PRNewswire) -

Auf der KubeCon + CloudNativeCon Europe 2024, die am 21. März in Paris stattfand, betonte Dennis Gu, leitender Architekt von Huawei Cloud, in einer Keynote mit dem Titel „Cloud Native x AI: Unleashing the Intelligent Era with Continuous Open Source Innovation" (Entfesselung des intelligenten Zeitalters mit kontinuierlicher Open-Source-Innovation), dass die Integration von nativen Cloud- und KI-Technologien entscheidend für die Transformation der Industrie ist. Huawei Cloud plant, weiterhin innovative Open-Source-Projekte zu entwickeln und mit Entwicklern zusammenzuarbeiten, um ein intelligentes Zeitalter zu schaffen.

KI stellt das Cloud-Native-Paradigma vor große Herausforderungen.

In den letzten Jahren haben native Cloud-Technologien herkömmliche IT-Systeme revolutioniert und den digitalen Fortschritt in Bereichen wie dem Internet und Behördendiensten beschleunigt. Cloud Native hat neue Möglichkeiten geschaffen, wie blitzschnelle Verkäufe und agile Abläufe, wie DevOps, durch Microservice Governance. Diese Veränderungen haben sich erheblich auf das Leben der Menschen ausgewirkt, und das rasche Wachstum und die weit verbreitete Anwendung von KI, einschließlich groß angelegter Modelle, sind zum Kernstück der Industrieintelligenz geworden.

Laut einer Epoch-Studie von 2023 hat sich die für grundlegende Modelle benötigte Rechenleistung alle 18 Monate verzehnfacht, was fünfmal schneller ist als die vom Moore'schen Gesetz für die allgemeine Rechenleistung vorhergesagte Wachstumsrate. Das Aufkommen dieses „neuen Moore'schen Gesetzes" aufgrund von KI und die Verbreitung groß angelegter KI-Modelle stellt eine Herausforderung für Cloud-native Technologien dar. In seiner Rede ging Dennis Gu auf die folgenden Kernpunkte ein:

- Eine niedrige durchschnittliche GPU/NPU-Auslastung treibt die Kosten für KI-Training und KI-Inferenz in die Höhe.
- Häufige Ausfälle von großen Modelltrainingsclustern verringern die Trainingseffizienz.
- Die komplexe Konfiguration von Großmodellen stellt hohe Anforderungen an die KI-Entwicklung.
- Der Einsatz groß angelegter KI-Inferenzen birgt das Risiko unvorhersehbarer Verzögerungen beim Zugriff der Endnutzer und birgt potenzielle Datenschutzprobleme.

Die Huawei Cloud AI-Innovation bietet Entwicklern Ideen zur Bewältigung von Herausforderungen.

Die zunehmende Größe von KI-Modellen erfordert mehr Rechenleistung, was Herausforderungen für Cloud-native Technologien mit sich bringt, aber auch Chancen für Innovationen in der Branche eröffnet. Dennis Gu berichtete über die KI-Innovationen von Huawei Cloud und bot Entwicklern einen Bezugspunkt für die Bewältigung der Herausforderungen.

Huawei Cloud verwendete KubeEdge, eine Cloud-native Edge-Computing-Plattform, um eine Multi-Roboter-Planungs- und Verwaltungsplattform zu erstellen. Mit dieser Plattform können die Benutzer der Plattform über Befehle in natürlicher Sprache mitteilen, was zu tun ist, und das System koordiniert mehrere Roboter am Rande der Stadt, um komplexe Aufgaben zu erfüllen. Das System ist mit einer dreiteiligen Architektur (Cloud, Edge Node und Roboter) konzipiert, um Herausforderungen wie das Verstehen natürlicher Sprache, die effiziente Planung und Verwaltung mehrerer Roboter und die typübergreifende Verwaltung des Roboterzugriffs zu bewältigen. Es verwendet umfangreiche Modelle zur Ausführung von Befehlen in natürlicher Sprache und führt Verkehrsvorhersagen, Aufgabenzuweisungen und Routenplanung durch. Die dreiteilige Architektur erhöht die Flexibilität der Roboterplattform, steigert die Effizienz des Managements um 25 %, reduziert die für die Systembereitstellung benötigte Zeit um 30 % und verkürzt die für die Bereitstellung neuer Roboter benötigte Zeit von Monaten auf Tage.

Bei einer führenden Plattform für die gemeinsame Nutzung von Inhalten in China, die über 100 Millionen aktive Nutzer pro Monat hat, sind die Empfehlungen auf der Homepage der wichtigste Service. Diese Funktion stützt sich auf ein Modell mit fast 100 Milliarden Parametern. Um dieses Modell zu trainieren, nutzt die Plattform einen Trainingscluster mit Tausenden von Rechenknoten, darunter Hunderte von Ps und Arbeitern für eine einzige Trainingsaufgabe. Es besteht also ein großer Bedarf an besserer Topologieplanung, hoher Leistung und hohem Durchsatz. Volcano, ein Open-Source-Projekt, verbessert die Unterstützung für KI- oder Machine-Learning-Workloads auf Kubernetes und bietet eine Reihe von Job-Management- und erweiterten Scheduling-Richtlinien. Volcano beinhaltet Algorithmen wie Topologie-bewusste Planung, Bin-Packing und Service Level Agreement (SLA)-bewusste Planung, was zu einer 20-prozentigen Verbesserung der Gesamttrainingsleistung und einer signifikanten Reduzierung der Betriebs- und Wartungskomplexität der Plattform führt.

Serverless AI steht an der Spitze der nativen Cloud-Entwicklung.

Viele Unternehmen und Entwickler stehen vor der Herausforderung, KI-Anwendungen effizient und zuverlässig zu betreiben und gleichzeitig die Betriebskosten zu minimieren. Huawei Cloud hat eine Lösung für dieses Problem entwickelt, indem es die wichtigsten Anforderungen von Cloud-nativen KI-Plattformen identifiziert und ein neues Konzept namens Serverless AI eingeführt hat.

Während seines Vortrags erläuterte Dennis Gu, dass Serverless AI komplexe Trainings- und Inferenzaufgaben durch intelligente

Empfehlungen für parallele Richtlinien vereinfacht und damit für Entwickler einfacher zu nutzen ist. Es umfasst auch eine adaptive Funktion zur automatischen GPU/NPU-Erweiterung, die die Ressourcenzuweisung dynamisch an Änderungen der Arbeitslast in Echtzeit anpasst und so eine effiziente Aufgabenausführung gewährleistet. Darüber hinaus gibt es in Serverless AI einen fehlerfreien GPU/NPU-Cluster, der die Entwickler von der Sorge befreit, dass Hardwarefehler die Dienste unterbrechen könnten. Vor allem aber ist Serverless AI mit den gängigen KI-Frameworks kompatibel, so dass Entwickler ihre vorhandenen KI-Tools und -Modelle problemlos integrieren können.

Serverless AI ist auch für Cloud-Service-Anbieter eine sehr wichtige Entwicklung. Serverless AI bietet zahlreiche Vorteile wie eine bessere GPU-/NPU-Auslastung, effizientere hybride Workloads für Training, Inferenz und Entwicklung sowie umweltfreundliches Computing durch bessere Energieeffizienz, sodass Sie Stromkosten sparen können. Darüber hinaus ermöglicht Serverless AI die gemeinsame Nutzung von GPU/NPU durch mehrere Mieter in unterschiedlichen Bereichen oder zu unterschiedlichen Zeiten, wodurch die Wiederverwendungsrate von Ressourcen verbessert wird. Der wichtigste Aspekt von Serverless AI ist die Fähigkeit, garantierte Dienstgüte (QoS) und SLAs sowohl für Trainings- als auch für Inferenzaufgaben bereitzustellen und so einen stabilen und hochwertigen Service zu gewährleisten.

Serverless AI verwendet eine flexible Ressourcenplanungsschicht, die auf einem virtualisierten Betriebssystem aufbaut. Diese Schicht kapselt wesentliche Funktionen von Anwendungsrahmenwerken in die Vermittlungsschicht für Anwendungsressourcen ein. Dennis Gu stellte die Referenzarchitektur für Serverless AI vor. Er ist der Meinung, dass dieses Architekturdesign es Serverless AI ermöglicht, automatisch große KI-Ressourcen zu betreiben. Dazu gehören die genaue Analyse von Ressourcennutzungsmustern, die gemeinsame Nutzung von Ressourcen aus heterogenen Hardwarepools und die Gewährleistung von Fehlertoleranz bei KI-Trainingsaufgaben durch GPU/NPU-Virtualisierung und Live-Lastmigration. Darüber hinaus verbessern die mehrdimensionale Planung und die adaptive elastische Skalierung die Ressourcenauslastung.

Auf dem Unterforum stellten technische Experten von Huawei Cloud fest, dass die KI- oder Machine-Learning-Workloads, die auf Kubernetes laufen, stetig zunehmen. Infolgedessen bauen zahlreiche Unternehmen Cloud-native KI-Plattformen über mehrere Kubernetes-Cluster auf, die über verschiedene Rechenzentren und eine Vielzahl von GPU-Typen verteilt sind. Karmada und Volcano können GPU-Workloads auf intelligente Weise über mehrere Cluster hinweg planen, die Fehlerübertragung unterstützen und Konsistenz und Effizienz innerhalb und zwischen Clustern sicherstellen. Sie können auch die Ressourcennutzung über das gesamte System und die QoS von Arbeitslasten mit unterschiedlichen Prioritäten ausgleichen, um die Herausforderungen der Verwaltung großer und heterogener GPU-Umgebungen zu bewältigen.

Karmada bietet ein sofortiges, zuverlässiges automatisches Anwendungsmanagement in Multi-Cloud- und Hybrid-Cloud-Szenarien. Eine wachsende Zahl von Anwendern nutzt Karmada, um anpassungsfähige und effektive Lösungen in Produktionsumgebungen zu schaffen. Karmada wurde 2023 offiziell zum CNCF-Inkubationsprojekt aufgewertet, und die Community freut sich auf weitere Partner und Entwickler, die sich anschließen.

Volcano Gang Scheduling ist eine Lösung für verteiltes KI-Training und Big-Data-Szenarien und löst das Problem des endlosen Wartens und der Blockade bei verteilten Trainingsaufgaben. Mit Task-Topologie und E/A-bewusster Planung wird die Übertragungsverzögerung verteilter Trainingsaufgaben minimiert, wodurch die Trainingsleistung um 31 % verbessert wird. Außerdem löst minResources die Ressourcenkonkurrenz zwischen dem Spark-Treiber und dem Executor in Szenarien mit hoher Parallelität auf, optimiert den Grad der Parallelität und verbessert die Leistung um 39,9 %.

Dennis Gu ist der Ansicht, dass der Schlüssel zur Verbesserung der KI-Produktivität in der Agilität nativer Cloud-Technologien und in der Innovation heterogener KI-Computing-Plattformen liegt. Huawei Cloud hat sich der Open-Source-Innovation verschrieben und will gemeinsam mit Branchenkollegen eine neue Ära der Intelligenz einläuten.

Foto - https://mma.prnewswire.com/media/2370741/Dennis_Gu_Chief_Architect_Huawei_Cloud.jpg

View original content:https://www.prnewswire.com/news-releases/huawei-cloud-auf-der-kubecon-eu-2024-entfesselung-des-intelligenten-zeitalters-mit-kontinuierlicher-open-source-innovation-302101619.html

Pressekontakt:

Lavanda Wang,
lavanda.wang@huawei.com

Diese Meldung kann unter https://www.presseportal.ch/de/pm/100090258/100917560 abgerufen werden.